CoRoLa Corpus

Contemporary Romanian Language Corpus

Priority Program of the Romanian Academy

2014-2017

extended for 3 more years

Dan Tufiș¹, Dan Cristea²
Romanian Academy

¹ Research Institut for Artificial Intelligence ”Mihai Drăgănescu” Bucharest
² Institut for Computer Science, Iași
Natural language in KBS

- Language communication is the fundamental way of human interaction
- Knowledge, accumulated by the humanity is mostly preserved and transmitted/disseminated in written or spoken form
- The last decades of tremendous advances in technology tend to replace much of the face-to-face human communication by technology-mediated interaction
- Major concern in Europe (also all around the world) on languages preservation in virtual spaces
META-NET White Paper Series: Press Release

At Least 21 European Languages in Danger of Digital Extinction

Good News and Bad News on the European Day of Languages

Most European languages are unlikely to survive in the digital age, a new study by Europe’s leading Language Technology experts warns. Assessing the level of support through language technology for 30 of the approximately 80 European languages, the experts conclude that digital support for 21 of the 30 languages investigated is “non-existent” or “weak” at best. The study “Europe’s Languages in the Digital Age” was carried out by META-NET, a European network of excellence that consists of 60 research centres in 34 countries, working on the technological foundations of multilingual Europe.

Europe must take action to prepare its languages for the digital age. They are a precious component of our cultural heritage and, as such, they deserve future-proofing. The European Day of Languages on September 26 recognises the importance of fostering and developing the rich linguistic and cultural heritage of our continent. The META-NET study shows that, in the digital age, multilingual Europe and its
### Major languages on the Internet

W3Techs estimated percentages of the top 10 million websites Nov.19, 2017

<table>
<thead>
<tr>
<th>Rank</th>
<th>Language</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>English</td>
<td>51.2%</td>
</tr>
<tr>
<td>2</td>
<td>Russian</td>
<td>6.7%</td>
</tr>
<tr>
<td>3-4</td>
<td>Japanese, German</td>
<td>5.6%</td>
</tr>
<tr>
<td>5</td>
<td>Spanish</td>
<td>5.1%</td>
</tr>
<tr>
<td>6</td>
<td>French</td>
<td>4.1%</td>
</tr>
<tr>
<td>7</td>
<td>Portuguese</td>
<td>2.6%</td>
</tr>
<tr>
<td>8</td>
<td>Italian</td>
<td>2.4%</td>
</tr>
<tr>
<td>9</td>
<td>Chinese</td>
<td>2.0%</td>
</tr>
<tr>
<td>10</td>
<td>Polish</td>
<td>1.7%</td>
</tr>
<tr>
<td>11</td>
<td>Turkish</td>
<td>1.6%</td>
</tr>
<tr>
<td>12</td>
<td>Persian</td>
<td>1.5%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>Language</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>Dutch, Flemish</td>
<td>1.4%</td>
</tr>
<tr>
<td>14-15</td>
<td>Korean, Czech</td>
<td>0.9%</td>
</tr>
<tr>
<td>16</td>
<td>Arabic</td>
<td>0.8%</td>
</tr>
<tr>
<td>17</td>
<td>Vietnamese</td>
<td>0.6%</td>
</tr>
<tr>
<td>18-21</td>
<td>Romanian, Indonesian, Greek, Swedish</td>
<td>0.5%</td>
</tr>
<tr>
<td>22</td>
<td>Hungarian</td>
<td>0.4%</td>
</tr>
<tr>
<td>23-26</td>
<td>Danish, Thai, Slovak, Finnish</td>
<td>0.3%</td>
</tr>
<tr>
<td>27-28</td>
<td>Bulgarian, Hebrew</td>
<td>0.2%</td>
</tr>
<tr>
<td>29-39</td>
<td>Lithuanian, Norwegian, Ukrainian, Croatian, Norwegian (B), Serbian, Catalan +Valencian, Slovenian, Latvian, Estonian, Tamil</td>
<td>0.1%</td>
</tr>
</tbody>
</table>

en.wikipedia.org/wiki/Languages_used_on_the_Internet
… and as of 8 September 2018

<table>
<thead>
<tr>
<th></th>
<th>Language</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>English</td>
<td>53.1%</td>
</tr>
<tr>
<td>2</td>
<td>German</td>
<td>6.3%</td>
</tr>
<tr>
<td>3</td>
<td>Russian</td>
<td>6.1%</td>
</tr>
<tr>
<td>4</td>
<td>Spanish</td>
<td>5.0%</td>
</tr>
<tr>
<td>5</td>
<td>French</td>
<td>4.2%</td>
</tr>
<tr>
<td>6</td>
<td>Japanese</td>
<td>3.6%</td>
</tr>
<tr>
<td>7</td>
<td>Portuguese</td>
<td>2.9%</td>
</tr>
<tr>
<td>8</td>
<td>Italian</td>
<td>2.4%</td>
</tr>
<tr>
<td>9</td>
<td>Persian</td>
<td>2.0%</td>
</tr>
<tr>
<td>10</td>
<td>Chinese</td>
<td>1.8%</td>
</tr>
<tr>
<td>11</td>
<td>Polish</td>
<td>1.8%</td>
</tr>
<tr>
<td>12</td>
<td>Dutch, Flemish</td>
<td>1.3%</td>
</tr>
<tr>
<td>13</td>
<td>Turkish</td>
<td>1.2%</td>
</tr>
<tr>
<td>14</td>
<td>Czech</td>
<td>1.0%</td>
</tr>
<tr>
<td>15</td>
<td>Korean</td>
<td>0.9%</td>
</tr>
<tr>
<td>16</td>
<td>Vietnamese</td>
<td>0.6%</td>
</tr>
<tr>
<td>17</td>
<td>Arabic</td>
<td>0.6%</td>
</tr>
<tr>
<td>18</td>
<td>Swedish</td>
<td>0.5%</td>
</tr>
<tr>
<td>19</td>
<td>Hungarian</td>
<td>0.5%</td>
</tr>
<tr>
<td>20</td>
<td>Greek</td>
<td>0.5%</td>
</tr>
<tr>
<td>21</td>
<td>Romanian</td>
<td>0.4%</td>
</tr>
<tr>
<td>22</td>
<td>Slovak</td>
<td>0.4%</td>
</tr>
<tr>
<td>23</td>
<td>Danish</td>
<td>0.3%</td>
</tr>
<tr>
<td>24</td>
<td>Indonesian</td>
<td>0.3%</td>
</tr>
<tr>
<td>25</td>
<td>Finnish</td>
<td>0.3%</td>
</tr>
<tr>
<td>26</td>
<td>Thai</td>
<td>0.2%</td>
</tr>
<tr>
<td>27</td>
<td>Bulgarian</td>
<td>0.2%</td>
</tr>
<tr>
<td>28</td>
<td>Ukrainian</td>
<td>0.2%</td>
</tr>
<tr>
<td>29</td>
<td>Hebrew</td>
<td>0.2%</td>
</tr>
<tr>
<td>30</td>
<td>Norwegian Bokmål</td>
<td>0.2%</td>
</tr>
<tr>
<td>31</td>
<td>Lithuanian</td>
<td>0.1%</td>
</tr>
<tr>
<td>32</td>
<td>Croatian</td>
<td>0.1%</td>
</tr>
<tr>
<td>33</td>
<td>Norwegian</td>
<td>0.1%</td>
</tr>
<tr>
<td>34</td>
<td>Serbian</td>
<td>0.1%</td>
</tr>
<tr>
<td>35</td>
<td>Catalan</td>
<td>0.1%</td>
</tr>
<tr>
<td>36</td>
<td>Slovenian</td>
<td>0.1%</td>
</tr>
<tr>
<td>37</td>
<td>Estonian</td>
<td>0.1%</td>
</tr>
<tr>
<td>38</td>
<td>Latvian</td>
<td>0.1%</td>
</tr>
<tr>
<td>39</td>
<td>Hindi</td>
<td>0.1%</td>
</tr>
</tbody>
</table>
What is CoRoLa?

- An electronic reference corpus of language samples for contemporary Romanian, written and spoken
  - 1945 – our days
- Original Romanian texts and native Romanians’ voices
  - but translations are allowed (metadata specifying this)
Building a corpus...
CoRoLa – achievement of a large corpus of texts

Corpus of Contemporary Romanian Language

Portal
Data Collection

 Paladin

• In an IPR-regulated world constructing a reference corpus requires clearing the property rights. Not everything on the Internet is IPR-free! Several international existing corpora might be liable for IPR infringement!

• Not the case for CoRoLa!

  • we concluded cooperation agreements with all the IPR owners of the documents we collected
  • some collected documents are IPR-free, but their aggregators were informed on our use
Getting primary data: Curator – Owner – Portal
Getting primary data: Curator – Owner– Portal
Getting primary data:
Curator – Owner – Portal
Getting primary data: Curator – Owner – Portal
Getting primary data:
Curator – Owner – Portal
Getting primary data:
Curator – Owner – Portal

Portalul COROLA
Curation chain:
Portal – Volunteers – Portal

CoRoLa Portal

Cleaning
1. Character
2. Headers
3. Footers
4. Formulas
5. Tables
6. Table of contents
7. Bibliography etc.

Metadata
1. Title
2. Author
3. Publication date
4. Source
5. Translator
6. Media
7. Style
8. Domain
9. ISSN/ISBN
CoDAP
CoRoLa Data cleaning and metadata Platform
(http://89.38.230.23/)
Before being subject to linguistic processing, the linguistically irrelevant parts of the data collected are purged out and text is cleaned:

- **TEXT**: boilerplate removal, character encoding, diacritics recovering, headers and footers eliminated, also pagination, tables and captions, formulas, footnotes, fragments written in other languages
- **SPEECH**: background music, publicity, pauses, fragments where transcriptions are not identical to the speech signal, inaudible sequences, etc.
Annotation chain: Portal

Layers:
- token
- part of speech
- morphology
- noun phrase
- syntax
- semantics
- ...

CoRoLa Portal

TTL pipeline

POST pipeline: indexing
Use of the corpus

Acces point: ICIA MD, Bucharest

Mirror: IIT-ARFI, Iași

Programs used in Computational Linguistics
What is CoRoLa intended for?

- Linguistic studies (both mono/multi-lingual)
- Language learning
- Building language technologies
What is CoRoLa intended for?

• Development of new and better technologies and applications for processing and using Romanian language in the digital world
Quantities...

TEXT

- Almost 400,000 files
- 1,26 billion tokens (words + punctuations) - as of 13 September
- Approx. 900,000,000 words
- Supposing 400 words per page ⇒ approx. 2,250,000 printed pages
- An average paperback book has about 250 pages per cm of thickness (cf. BNC-xml edition) ⇒ the whole corpus, printed in small type on thin paper would take up about 90 meters of shelf space
- XML annotations multiplies the volume by a factor of 8 (let alone the metadata)
- Surface unique forms: > 3 million
  - frequency higher than 50: 198,800 words (obsolete)
- Lemmas unique forms: 2,346,546, out of which lowercase lemmas: 2,136,391
  - frequency higher than 50: 121,091 lemmas (obsolete)
Values are orientative, continuously updated

- Two categories of oral (speech-data) documents:
  - **Collected data**
    - more than 300 hours of recorded speech, not entirely processed
    - IPR-cleared
  - **Processed data**
    - transcribed and aligned speech data: 71.5 hours (already ready to use)
    - 137,351 spoken words
• Major functional styles:
  – Imaginative
  – Journalistic
  – Administrative
  – BlogPost
  – Science
  – Law
  – Memoires
Domains

- 4 major domains (plus one place-holder, Other):
  - Art&Culture; Nature; Science; Society

- Further refined into 70 sub-domains:
Accessing CoRoLa

- Open access to any interested user, via three available interfaces: NL2CQP (100%), OCQP (55%), KorAP (100%)

- Downloading a limited volume of data (the results of search queries)
- Any spotted errors, or suggestions for improvement may be submitted to the development teams via the access interfaces
Concordances
(KWIC – Key Word In Context)

http://Corola.Racai.RO/
Who are behind CoRoLa
Gratefully acknowledged on the CoRoLa site

- **Data providers:**
  Media, Publishing houses, Public institutions, Bloggers

- **Volunteers:**
  Linguists, PhD Students, MSc Students

- **International partnerships**
  IDS-Mannheim+Univ. Buc+ICIA+ITI

- **International support**
  “Alexander von Humboldt” Foundation

- **Development teams:**
  ICIA-MD București, ITI-Iași, IDS-Mannheim
Who are behind CoRoLa
Gratefully acknowledged on the CoRoLa site

Media

S.C. Press Media Electronic S.R.L. (DCNEWS) (București)
Revista România literară (București)
Uniunea Compozitorilor și Muzicologilor din România (București)
Societatea Română de Radiodifuziune (București)
Revista Balcanii și Europa (București)
Gazeta de Artă Politică (București)
Ziarul Agenda (Timișoara)
Revista Medicală Română (București)

SC Dorithea Media SRL – RomanTV (Roman)
Stil Media SRL - Radio VIVA fm (Iași)
Radio România Iași (Iași)
Radio Universitas (Iași)
Revista Cronica Timpului (București)
Revista UZP (București)
Candela de Montreal (Canada)
Destine Literare (Canada)
Presa Online
INFOIași
Timpul
Who are behind CoRoLa
Gratefully acknowledged on the CoRoLa site

Publishing Houses

Editura Academiei Române
Editura Economică
Editura Humanitas
Editura POLIROM
Editura Simetria
Editura Universității București
Editura Papirus Media
Editura Serafica

Fundația “Română 2000” SC AMO-PRODPRESS SRL
Casa Editorială Demiurg
Editura Babel
Editura Adenium
Editura PIM
Editura Gama
Editura Doxologia
Editura Institutul European
Editura ARS Longa
Who are behind CoRoLa
Gratefully acknowledged on the CoRoLa site

Public Institutions

• Universitatea ”A.I. Cuza” Iași
• Institutul de Filologie Română
  ”Alexandru Philippide”, Iași
• Institutul de Lingvistică
  ”Iorgu Iordan – Al. Rosetti”, București
• Universitatea București
• Federația Română de Diabet, Nutriție și Boli Metabolice
• Colegiul Național Unirea (Focșani, jud. Vrancea)
• Teatrul Național Târgu-Mureș
• Teatrul Național Cluj-Napoca
Who are behind CoRoLa
Gratefully acknowledged on the CoRoLa site

Bloggers & associations

- Simona Tache
- Dragoș Bucurenci
- Teodora Forăscu
- Irina Șubredu
- Andreea Ignat
- Blog de părinți
- Belva
- Iași4U – Orașul tău online
- Elena Filip
- Sebastian Bârgău
- Ramona Cervenciuc

http://printreranduri.eu
http://www.scriitoriiromani.com/ (Asociația Scriitorilor Romani din Canada)
http://uzp.org.ro/ (Uniunea Ziariștilor Profesioniști din România)
http://www.zilesinopti.ro/bucuresti
http://www.opiniastudenteasca.ro/
# Who are behind CoRoLa

Gratefully acknowledged on the CoRoLa site

## Volunteers

<table>
<thead>
<tr>
<th>Cosmin-Constantin Andrei</th>
<th>Ioana Curcă</th>
<th>Andrei Anton Popescu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ioan Baciu</td>
<td>Zeno Fodor</td>
<td>Gabriela-Diana Purice</td>
</tr>
<tr>
<td>Maria-Tereza Barnea</td>
<td>Roxana Hrăniciuc</td>
<td>Cristian Radu</td>
</tr>
<tr>
<td>Lavinia Maria Băian</td>
<td>Alexandru Iliescu</td>
<td>Francesca Sacaloș</td>
</tr>
<tr>
<td>Roxana Lumița Belciug</td>
<td>Alina Leonte</td>
<td>Alexandru Sălăvăstru</td>
</tr>
<tr>
<td>Mihaela Beța</td>
<td>Corneliu Leu</td>
<td>Diana–Alexandra Soponaru</td>
</tr>
<tr>
<td>Mădălina Maria Bîrzu</td>
<td>Andreea Evelina Levițchi</td>
<td>Amalia Maria Tanasă</td>
</tr>
<tr>
<td>Ionuț Răducu Căciulă</td>
<td>Andrei Cristian Luca</td>
<td>Ana-Maria Timofciuc</td>
</tr>
<tr>
<td>Lumița Cărăuşu</td>
<td>Ana-Maria Lungu</td>
<td>Gabriela Torică</td>
</tr>
<tr>
<td>Adriana Chițac</td>
<td>Adriana Moroșan</td>
<td>Andreea Țigănescu</td>
</tr>
<tr>
<td>Adina Cirobatariu</td>
<td>Monica Pălimăriu</td>
<td>Adina Zaharia</td>
</tr>
<tr>
<td>Cătălina Cojocari</td>
<td>Anca Pânzăriu</td>
<td>Evelina Zăporojanu</td>
</tr>
<tr>
<td>Ana-Maria Crețu</td>
<td>Liviu Petcu</td>
<td></td>
</tr>
</tbody>
</table>
International Cooperation

A huge boost in the development of CoRoLa: partnership with Institute for German Language – Mannheim (IDS), the developer of the German language corpus, the largest in the world (>30 billion tokens) and of the powerful corpus management platform (KoRaP) which we adopted.

DRuKoLa (2016-2018): made possible by the funding from “Alexander von Humboldt” Foundation
EuReKo: a crystallization structure for a federated reference corpus
Who are behind CoRoLa
Gratefully acknowledged on the CoRoLa site

Directly involved people

Two research institutes of the Romanian Academy:

ICIA-MD, București:
Verginica Barbu-Mititelu, Tiberiu Boros, Eric Curea, Ştefan Dumitrescu, Radu Ion, Elena Irimia, Maria Mitrofan, Vasile Păiş, Dan Tufiş.

IIT-ARFI, Iaşi:
Cecilia Bolea, Dan Cristea, Daniela Gîfu, Alex Moruz, Laura Pistol, Mihaela Plămădă-Onofrei, Andrei Scutelnicu

University of Bucharest:
Ruxandra Cosma
Corpus computațional de referință pentru limba română contemporană

Căutați un cuvânt în CoRoLa!
Launches and dissemination

- Already done:
  - Bucharest, Aula of the Romanian Academy Library, 14 December 2017
  - Chişinău, Science Academy of Republic of Moldova, 25 May 2018
  - Iaşi, Institute of Linguistics of the Romanian Academy, 30 May 2018

- Other possible dissemination places:
  - Cluj-Napoca, Timișoara, etc.

- Orchestrate the use of CoRoLa in faculties of Letters

- Collect reactions from users and to enhance access and content (through the online mailing service)
Is this an end?

- By no means! A reference corpus is a long-time enterprise!
- Still much work to do…
Cleaning should go on...

WE WANT AN ERROR-FREE CORPUS!

Sources of bugs...
- invalid codes for diachritics
- incompatibilities between the TTL annotations and the KorAP interface
- errors in metadata
- alignment errors between speech and text
- etc.

Filtering bugs
- programmatically
- manually
  => using the errors reporting service
A continuous development activity

- Included in the annual plans of IIT and ICIA
- Develop more annotation layers
  - NPs, VPs
  - syntax
  - mix syntax with semantics

Thanks to Cătăлина Mărănduc, PhD report, Dec. 2017
Look into the future

- Acquire more textual and speech data
  - get new providers
  - balance domains and styles
  - organize continuous crawling and fetching services for online data => automatic acquisition (the Mannheim model)
- Through permanent development: synchronic with the language use
  - “contemporary Romanian” should be appropriate now and in 50 years from now on
  - only keeping the present day data it will become an obsolete corpus, still not a diachronic one
Look into the past: towards a diachronic corpus

A Cyrillic-Romanian OCR-transcription pipeline: CyRo – a project in the second phase of evaluation, ePSALM – application upload: 30 Sept.

- semi-automatic acquisition of 16-19 century Romanian documents in manuscript, semi-uncial and printed form
- inference of paradigmatic morphology of old Romanian

- STEPS: documents ⇒ scanned ⇒ OCRed ⇒ transcribed ⇒ manually revised ⇒ POS-tagged, etc. ⇒ manually revised ⇒ included in the diachronic corpus ⇒ interfaced dedicated to humanist researchers
CyRo, ePSALT – deciphering old Cyrillic Romanian

Thanks to Gabriela Haja, the CyRo project proposal, July 2017
Improve access and utility

- Augment the corpus utility for *researchers of language* and *occasional users*
  - mark and export (list of) occurrences and sub-corpora
  - use syntactic and semantic constrains
    - enhance the Polycarp and NL2CQP query languages with more types of labels
- Allow the machine to access the corpus: *services*
  - open it for computational linguistics tasks
  - connect applications to CoRoLa
  - use *Linguistically Linked Open Data* methods to link CoRoLa onto other linguistic resources ➜ see tomorrow speech…
  - look at CoRoLa as big textual data ➜ see Dan Tufiş’s speech
Thank you!